Algorithmic Fairness from the lens of Causality and Information Theory

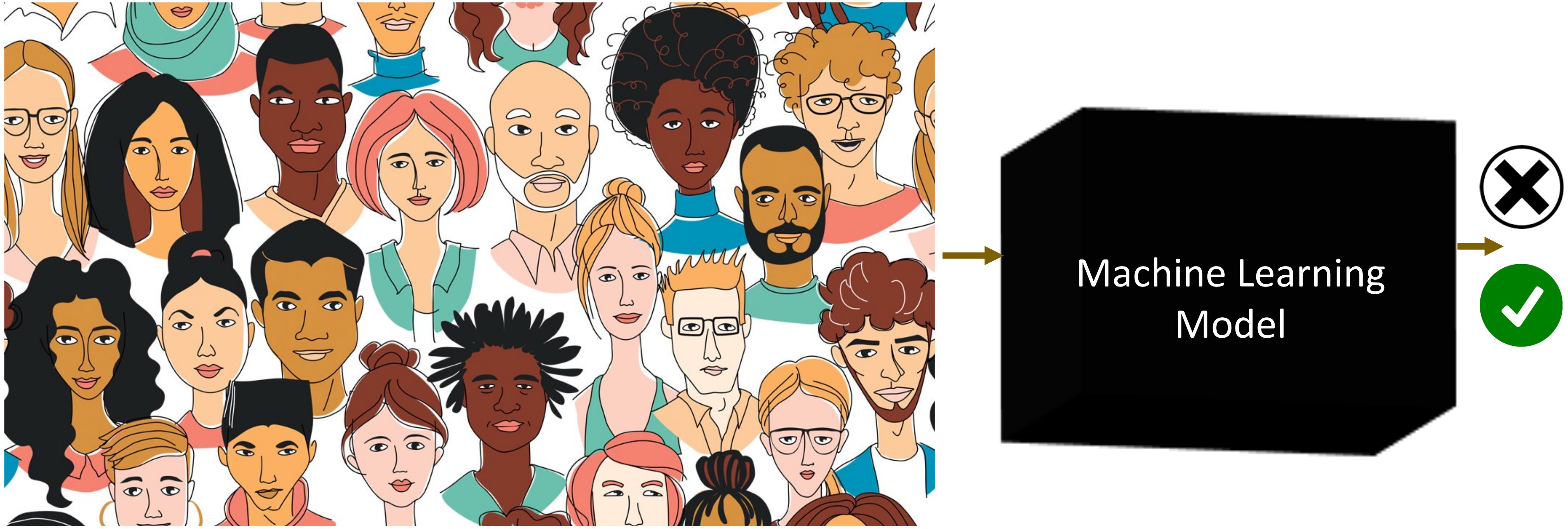
Sanghamitra Dutta

Workshop on Algorithmic Aspects of Causal Inference

Simons Institute

Carnegie Mellon University

Motivation: Machine Learning in High-Stakes Applications



HIRING EDUCATION LENDING HEALTHCARE

Motivation: Machine Learning in High-Stakes Applications

CNBC
MARKETS BUSINESS INVESTING TECH POLITICS CNBC TV WATCHLIST PRO

Amazon scraps a secret A.I. recruiting tool that showed bias against women

CNN BUSINESS
Markets Tech Media Success Perspectives Videos

Facebook settles lawsuits alleging discriminatory ads

How to identify/explain the sources of disparity in machine learning models?

How to identify/explain sources of disparity in machine learning models?

Example: Hiring a Software Engineer for a Safety-Critical Application

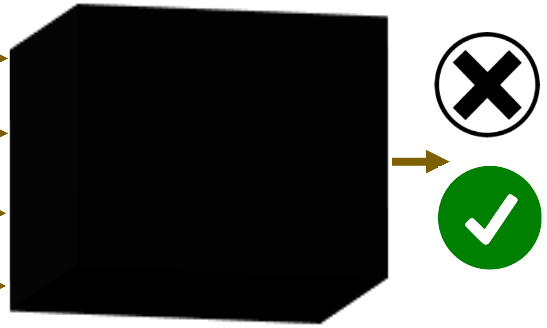
Inputs *Output*

Coding Test Score

Aptitude Test Score

All-Subject Grade

Gender, Race



Legal Motivation:
"Disparate Impact"

Title VII of Civil Rights Act: Disparate impact may be exempt if justified by "occupational necessity"

Coding Test may be critical

Weightlifting may be critical

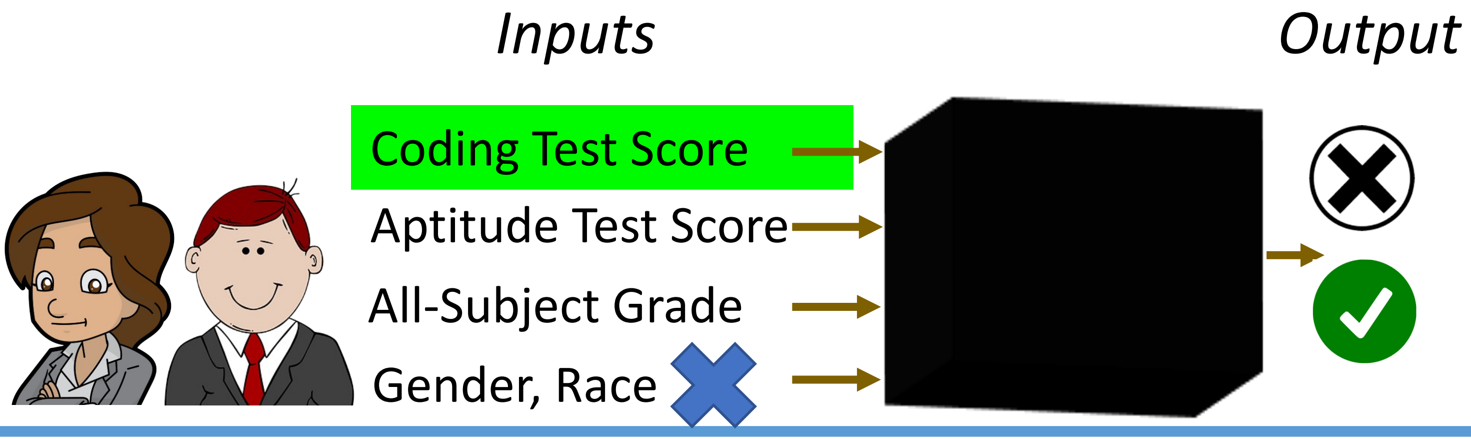
Griggs *v.* Duke Power Co.'71

Aptitude Test may not be critical

[Dwork et al.'12] [Grover'96][Barocas & Selbst'16][Feldman et al.'15]

How to identify/explain sources of disparity in machine learning models?

Example: Hiring a Software Engineer for a Safety-Critical Application



Q: Given a choice of critical features,
how do we say if the disparity is **exempt** or **non-exempt**?

Main Contribution:

A <u>systematic</u> measure of **non-exempt disparity**: bias not justified by critical features

*[**Dutta**, Venkatesh, Mardziel, Datta, Grover, AAAI'20; IEEE Trans. Information Theory'21]*

Algorithmic Fairness: A Growing Field of Research

Observational measures:

- **Statistical parity** [Agarwal et al.'18] [Calmon et al.'17]
- **Equalized odds** [Hardt et al.'17][Angwin et al.'16]
- **Predictive Parity** [Dieterich et al.'16][Chouldechova'16]
- **Proxy-Use** [Datta et al.'17] [Yeom et al.'18]
- **Disparate Impact** [Feldman et al.'15]
- **Subgroup/Conditional Fairness** [Kearns et al.'17][Corbett-Davis et al.'17][Kamiran et al.'12]

Causal measures: [Kusner et al.'17][Kilbertus et al.'17][Coston et al. '20][Zhang et al.'18][Nabi et al.'18]

Individual Fairness: [Dwork et al.'12]

Broad Perspective on Fairness: [Barocas & Hardt'17][Chouldechova & Roth'20][Varshney'19]

Other Related Works: [Galhotra et al.'20][Lipton et al.'17][Zafar et al.'17][Zemel et al.'13][Kamishima et al.'12] [Corbett-Davies et al.'17][Kamiran et al.'12][Salimi et al.'19] …. and many others

Quantify **non-exempt disparity** using
"Partial Information Decomposition" + Causality

Outline

How to identify/explain the sources of disparity in machine learning models?

Find a measure of non-exempt disparity

[AAAI 2020; IEEE Trans. Info Theory 2021]

Beyond Fairness: Application to Social Media & Filter Bubbles

[BIAS@ECIR 2021]

Perspectives on Accuracy-Fairness Tradeoffs

[ICML 2020] [NeurIPS 2021]

Connections with Explainability

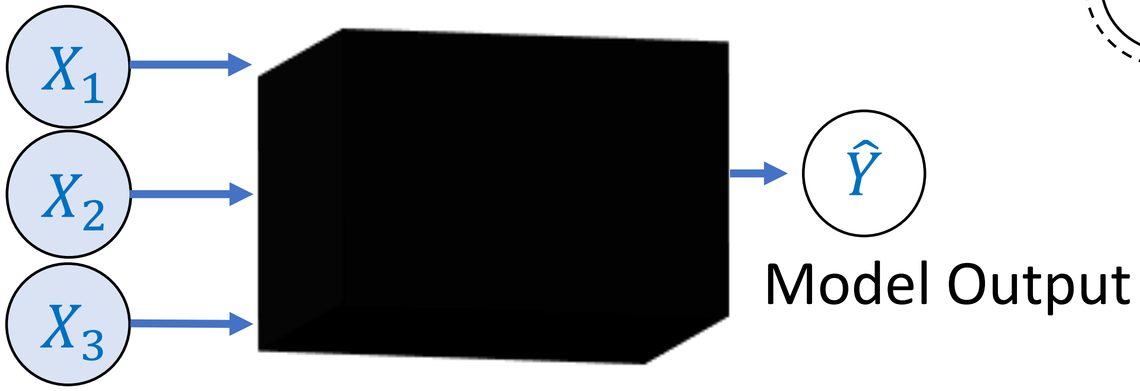
[Workshop@AAAI 2022]

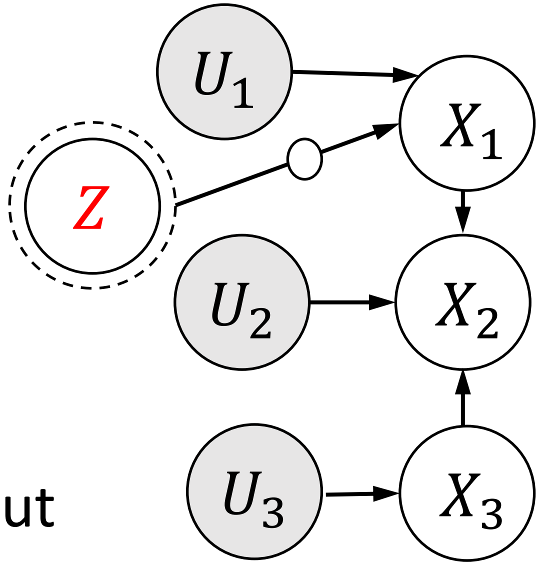
Problem Statement

Z: Protected Attribute, e.g., Gender, Race, etc.

Critical Features $X_c = X_1$

Non-Critical/General Features: $X_g = (X_2, X_3)$





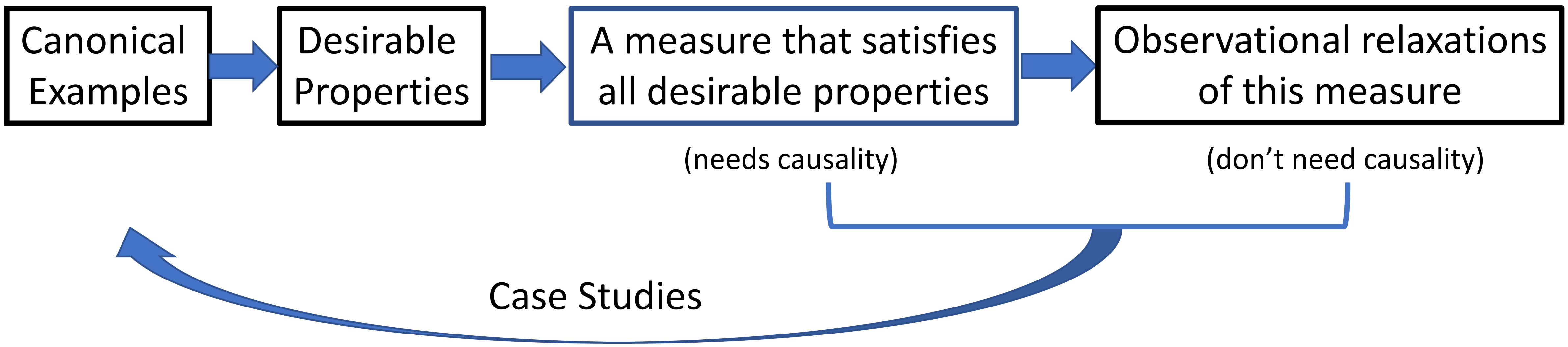
Given a choice of critical features X_c, what is a good measure of **non-exempt disparity** (M): bias that cannot be justified by critical features X_c?

Auditing: Compute $\underbrace{\text{M}}_{\text{non-exempt disparity}}$ on trained models

What is a good measure of **non-exempt disparity** (M)?

Training: $\min_{h(.)} \text{Loss}(Y, \hat{Y}) + \lambda \underbrace{\text{M}}_{\text{non-exempt disparity}}$ where $\hat{Y} = h(X)$

An axiomatic approach to arrive at a measure of non-exempt disparity



Canonical examples distill "paradoxical" scenarios
where a candidate measure may fail

Pros & cons of several candidate measures

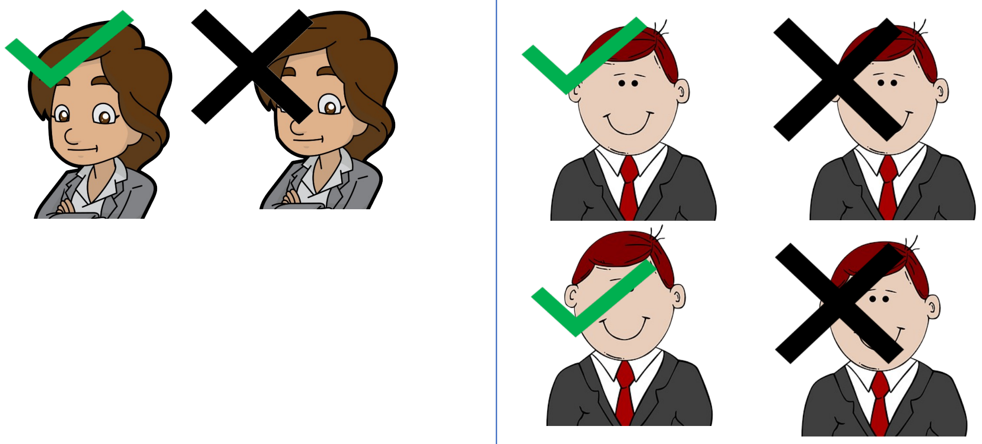
Popular Definitions: Statistical Parity and Equalized Odds & Their Pros and Cons

Popular Definition: Statistical Parity

$$\Pr(\hat{Y} = y | Z = 0) = \Pr(\hat{Y} = y | Z = 1)$$

Z: Gender (0/1), $\hat{Y}$: Model Output (✓ / ✗)

Women $(Z = 0)$	Men $(Z = 1)$



$\Pr(\hat{Y} =$ ✓ $) = 1/2$ $\Pr(\hat{Y} =$ ✓ $) = 1/2$

Model is fair if
$\hat{Y}$ is INDEPENDENT of Z

Information-theoretic measure of statistical **disparity**: $\mathrm{M} = I(Z; \hat{Y})$

$$I(Z; \hat{Y}) = \sum_{z,y} p(z, y) \log \frac{p(z, y)}{p(z)p(y)}$$
$$= D_{KL}\left(p_{(Z,\hat{Y})} || p_Z p_{\hat{Y}}\right)$$

Statistical Dependency

[Agarwal et al.’17][Zliobaite et al.’15]
Some Criticisms: [Zemel et. al.’13][Datta et. al.’17][Kusner et. al.’17][Hardt et. al.’16]

Criticism: Statistical Parity may disregard critical necessities

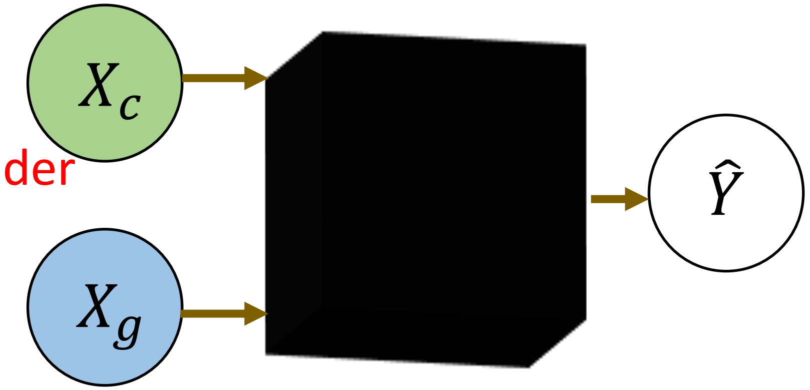
Accept applicants who may not meet critical necessities

Software Engineer for a Safety-Critical Application

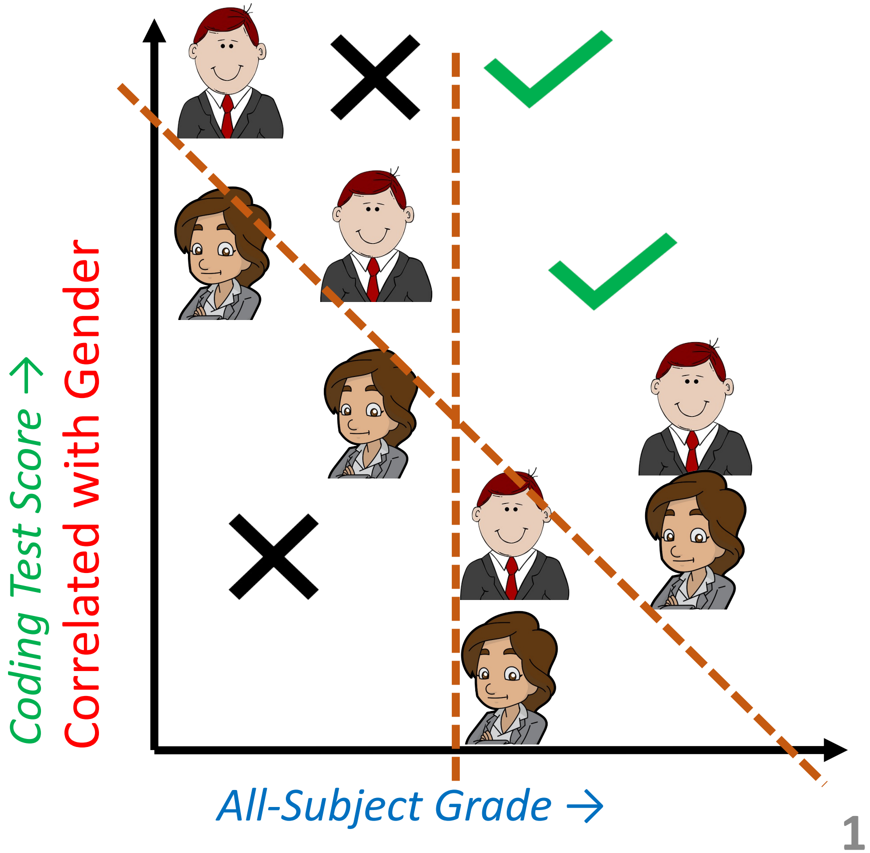
Critical Feature:
Coding Test Score

Correlated with Gender

General Feature:
All-Subject Grade



Model may significantly reduce emphasis on critical feature *Coding Test Score*



Popular Definition: Equalized Odds

$$\Pr(\hat{Y} = y | Z = 0, Y = y') = \Pr(\hat{Y} = y | Z = 1, Y = y')$$

Z: Gender (0/1), $\hat{Y}$: Model Output (✓/✗), Y : True Labels (✓/✗)

Model is fair if
$\hat{Y}$ is INDEPENDENT of Z **conditioned on** Y (True Labels)

Perfect classifier $\hat{Y} = Y$ satisfies Equalized Odds

[Hardt et al.'16]
Some Criticisms: [Hinnefeld'18][Yeom et al.'18][Barocas & Selbst'16]

Criticism: Equalized Odds regards past labels as infallible

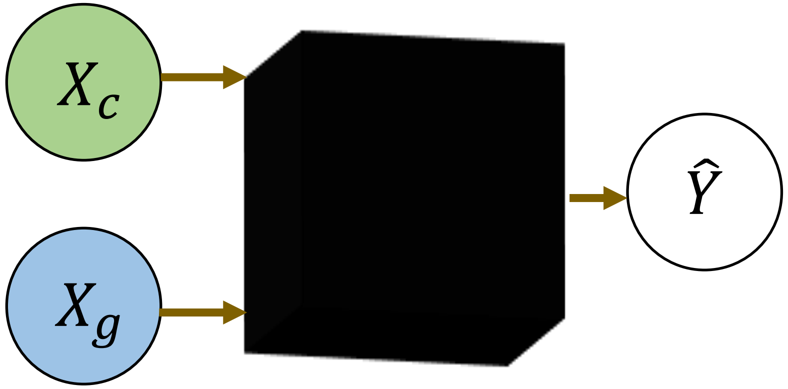
Agreement with historic labels propagates bias
(even for perfect classifiers that satisfy equalized odds)

Software Engineer for a Safety-Critical Application

Critical Feature:
Coding Test Score

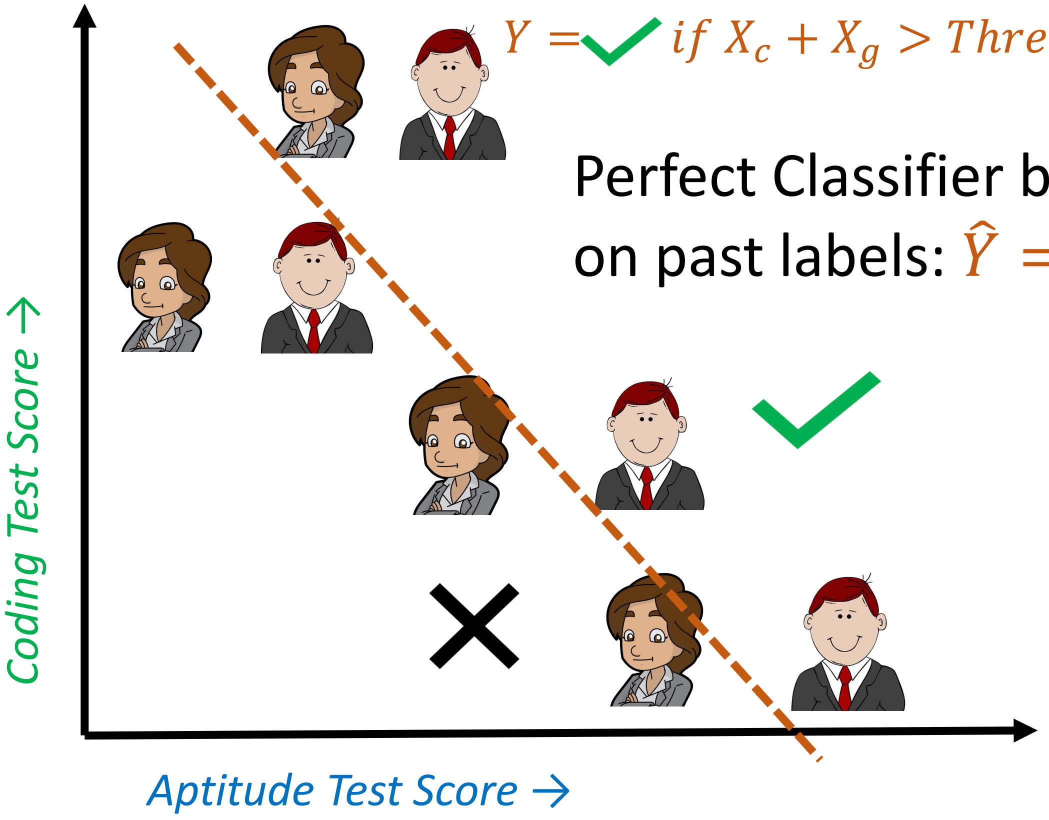
General Feature:
Aptitude Test Score

Correlated with Gender



$Y =$ ✓ $if\ X_c + X_g > Threshold$

Perfect Classifier based on past labels: $\hat{Y} = Y$



Correlated with Gender

Even a perfect classifier $\hat{Y} = Y$ may be illegal:

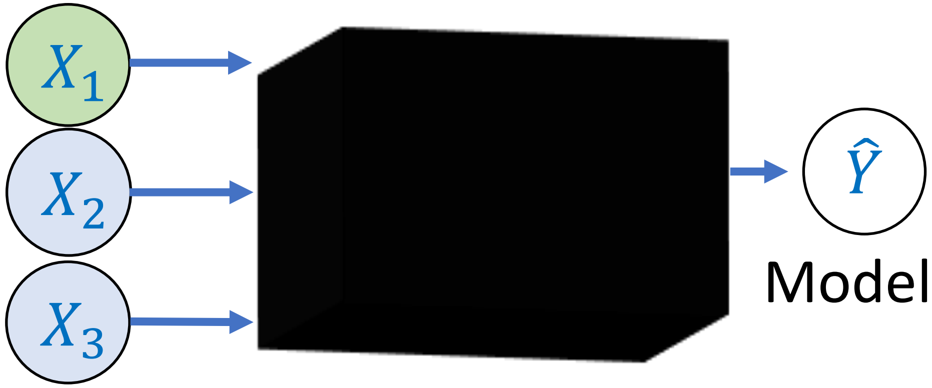
Aptitude Test Score not critical

E.g.,[Griggs *v.* Duke Power Co. '71]

Middle Ground between Statistical Parity and Equalized Odds using Domain Knowledge

Critical Features $X_c = X_1$

Non-Critical/General Features: $X_g = (X_2, X_3)$



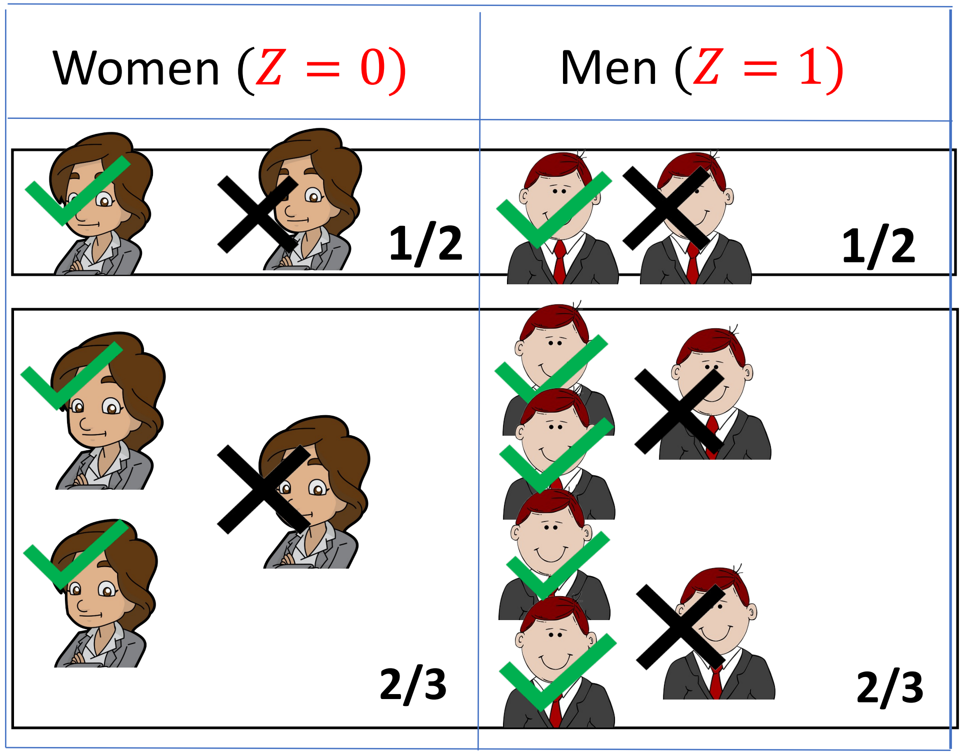
Model Output

What is a good measure of **non-exempt disparity** (M)?

Candidate Measure 1: Conditional Dependence $\mathrm{M} = I(Z;\hat{Y}|X_c)$

$$\Pr(\hat{Y}=y|Z=0, X_c=x_c) = \Pr(\hat{Y}=y|Z=1, X_c=x_c)$$

Z: Gender (0/1), $\hat{Y}$: Model Output (✓/✗)



Coding Test Score
$X_c = 1$

Coding Test Score
$X_c = 2$

$\Pr(\hat{Y} = ✓) = 3/5$ $\Pr(\hat{Y} = ✓) = 5/8$

Model is fair if $\hat{Y}$ is INDEPENDENT of Z **conditioned on** X_c

Our information-theoretic measure: $\mathrm{M} = I(Z;\hat{Y}|X_c)$

Inspired from [Corbett-Davies et al.'17][Kamiran et al.'12][Kilbertus et al.'17]

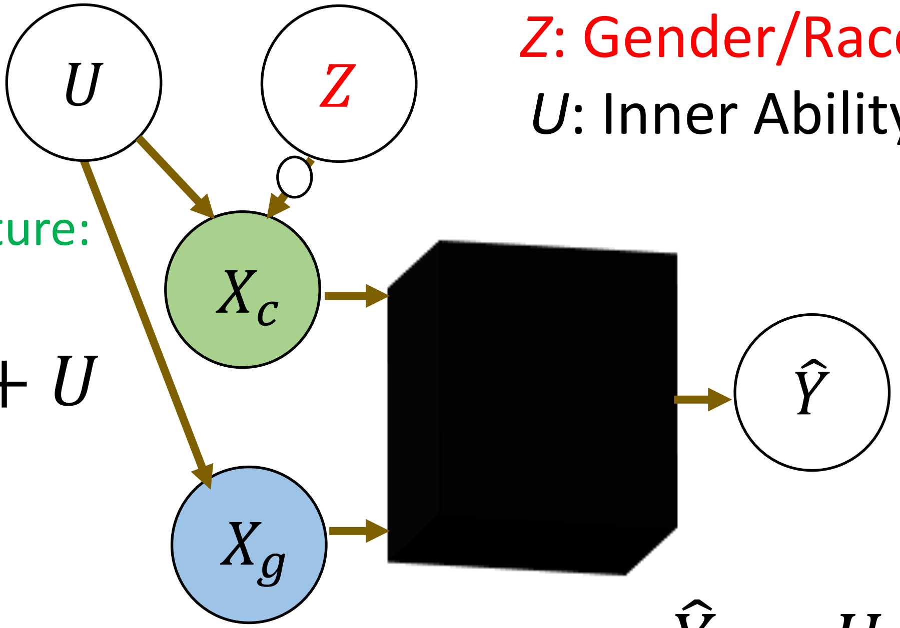
Our Key Observation:

Conditional Dependence can sometimes falsely detect bias (misleading dependencies) even when a model is “causally” fair

[Dutta et al. AAAI ’20; IEEE Trans. IT ‘21]

Conditional Dependence can sometimes falsely detect bias (misleading dependencies) even when a model is “causally” fair

Example: Causally fair model



Z: Gender/Race

U: Inner Ability

Correlated Critical Feature:
Coding Test Score

$$X_c = Z + U$$

General Feature:
Aptitude Test Score

$$X_g = U$$

$$\hat{Y} = U$$

Causally fair: $\hat{Y}$ doesn’t vary with Z after fixing inner ability U

$$Z \perp \hat{Y} | X_c$$

$$\mathrm{M} = I\big(Z; \hat{Y} | X_c\big) > 0$$

(falsely detects bias)

$$\Pr(\hat{Y} = y | Z = 0, X_c = x_c)$$
$$\neq \Pr(\hat{Y} = y | Z = 1, X_c = x_c)$$

Desirable Property 1:

A measure of non-exempt disparity M should be 0 if model is “causally” fair

[Dutta et al. AAAI ’20; IEEE Trans. IT ‘21]

Reference for Definition of Causal Fairness: [Kusner et al.’17]

Conditional Mutual Information does not satisfy
our “causal fairness” property

Conditional Mutual Information decomposes as:

Unique Information + Synergistic Information

satisfies our “causal fairness” property & some others

Theory of Partial Information Decomposition [Williams & Beer,’10] … [Bertschinger et al.’14]

Candidate Measure 2:
Unique Information $\mathrm{M} = Uniq(Z{:}\hat{Y}|X_c)$

Critical Feature: $X_c = Z + U$	Critical Feature: $X_c = U$
Output: $\hat{Y} = U$	Output: $\hat{Y} = Z + U$
Output $\hat{Y}$ has no information about gender Z	Output $\hat{Y}$ has some information about gender Z not in critical feature X_c

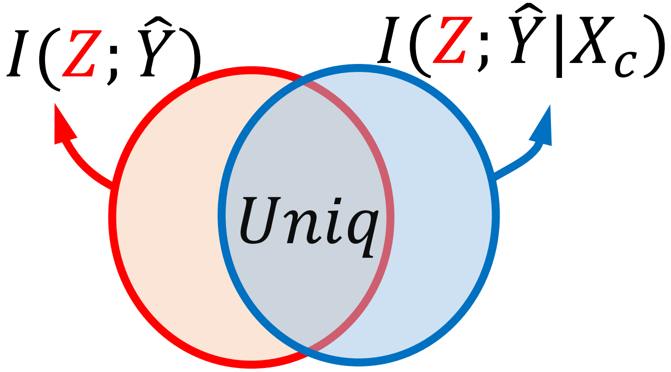
Z: Gender, Race

U: Inner Ability

$I(Z;\hat{Y}|X_c)$ is same for both these examples

Desirable Property 2: Distinguish between these two cases

$$Uniq(Z{:}\hat{Y}|X_c) = \min_{Q(Z,\hat{Y},\widetilde{X_c})} I(Z;\hat{Y}|\widetilde{X_c}) \text{ s.t. } Q(Z,\widetilde{X_c}) = P(Z,X_c)$$



$Uniq(Z{:}\hat{Y}|X_c)$ satisfies Property 1 (causal fairness) & Property 2

[Dutta et al. AAAI '20; IEEE Trans. IT '21]

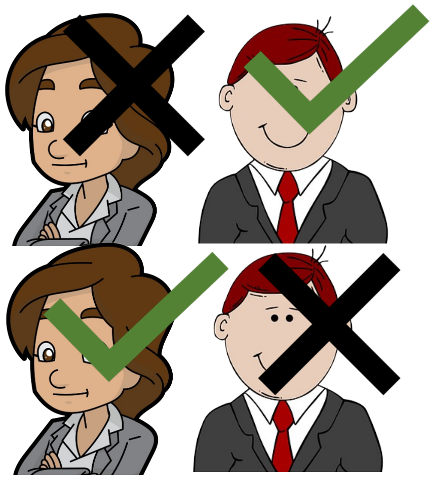
More nuanced issue that $Uniq(Z{:}\hat{Y}|X_c)$ does not address: "Masking"

Example: Masking in Hiring ADs

Inner Ability

$U = 1$

$U = 0$



Z: Gender, Race

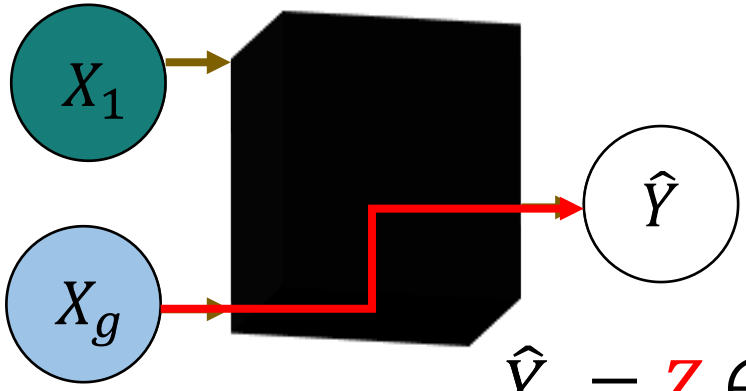
U: Inner Ability

Critical/General Feature:

$$X_1 = U$$

Correlated General Feature:

$$X_g = Z$$



$$\hat{Y} = Z \oplus U$$

Statistical disparity

$$I(Z;\hat{Y}) = 0$$

But not causally fair

Desirable Property 3: M should be non-zero in this example, detecting masking

One causal measure that satisfies all desirable properties

Theorem: Our proposed measure of **non-exempt disparity**, given by,

$$\mathrm{M}^{*} = \min_{U_a} Uniq\left((U_a, Z) : (\hat{Y}, U_b) | X_c\right)$$

satisfies our six desirable properties. Here U is the set of all latent random variables and $U_a = U \backslash U_b$.

- Property of Causal Fairness
- Property of Non-Exempt Visible Disparity
- Property of Non-Exempt Masked Disparity
- Property of Complete Exemption if $X_c = X$
- Property of Monotonicity with X_c
- Property of Zero Exemption if $X_c = \phi$

CAUSAL than CASUAL

- Benchmark for observational measures (pros/cons)
- Observational $Uniq(Z: \hat{Y} | X_c)$ is good enough except for masking

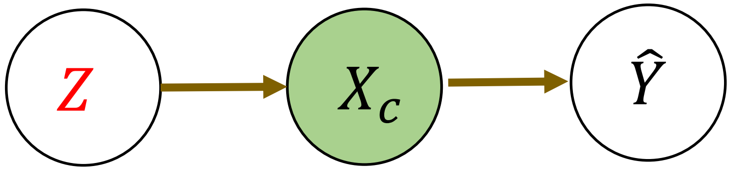
$$Uniq(Z: \hat{Y} | X_c) \underbrace{\leq}_{\text{“Masked”}} \min_{U_a} Uniq\left((U_a, Z) : (\hat{Y}, U_b) | X_c\right) \quad \text{for any set } U_a = U \backslash U_b$$

[Dutta et al. AAAI ’20; IEEE Trans. IT ‘21]

Some intuition on our proposed measure from causality

Is **non-exempt disparity** M=0 if all causal paths from Z to $\hat{Y}$ pass through X_c?

Example: Disparity Amplification

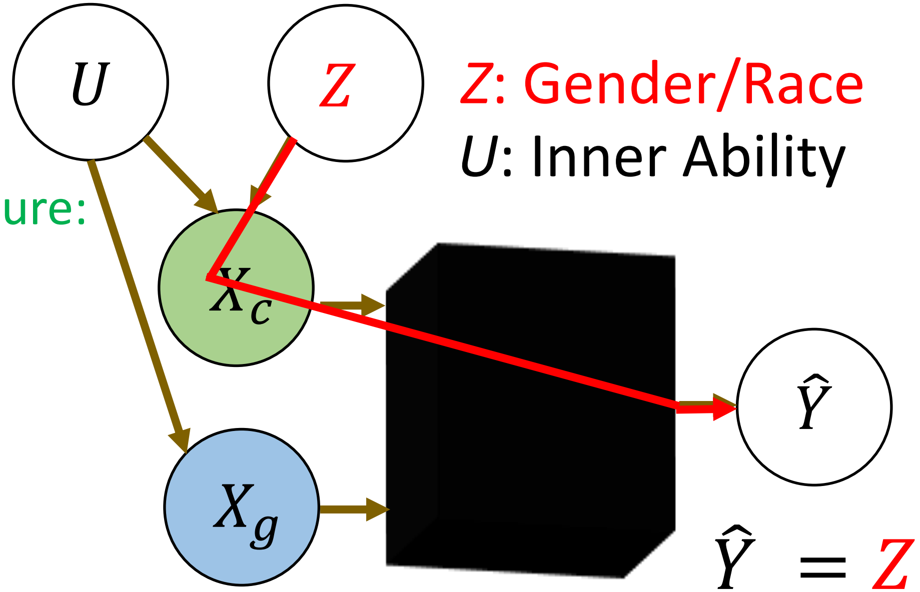


Correlated Critical Feature: *Coding Test Score*

$$X_c = Z + U$$

General Feature: *Aptitude Test Score*

$$X_g = U$$



Z: Gender/Race
U: Inner Ability

$$\hat{Y} = Z$$

Seemingly less-biased features mix to produce heavily-biased output $\hat{Y}$

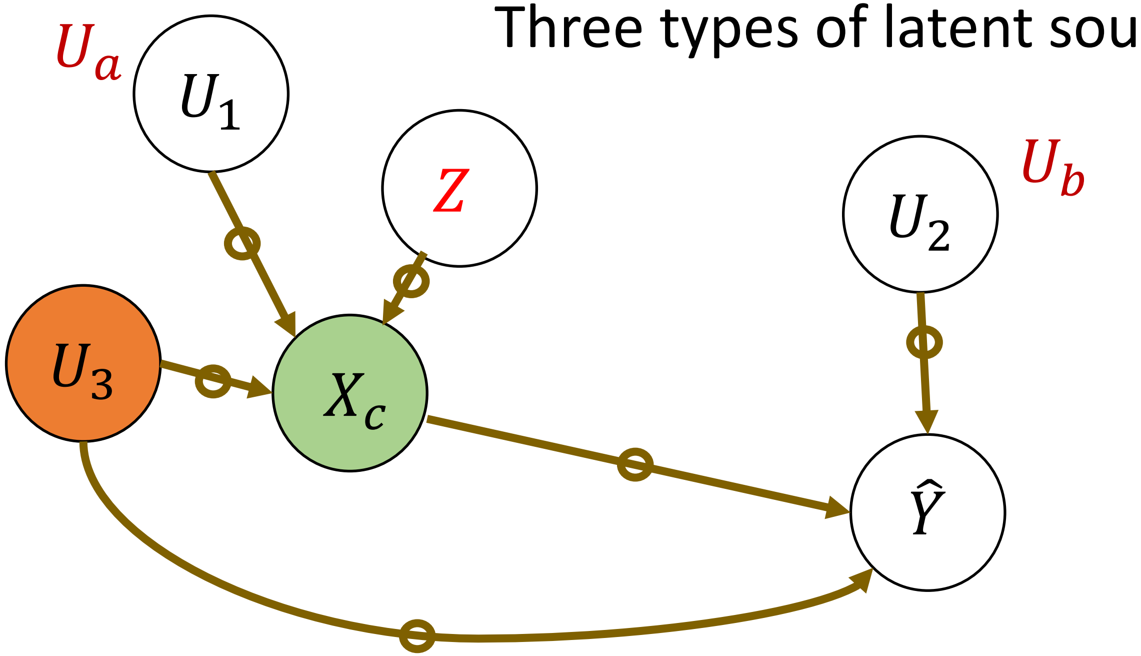
All causal paths from Z to $\hat{Y}$ pass through X_c

But U has confounding effects on X_c and $\hat{Y}$

Some intuition on our proposed measure from causality

Is **non-exempt disparity** M=0 if all causal paths from Z to $\hat{Y}$ pass through X_c?

Three types of latent source variables U

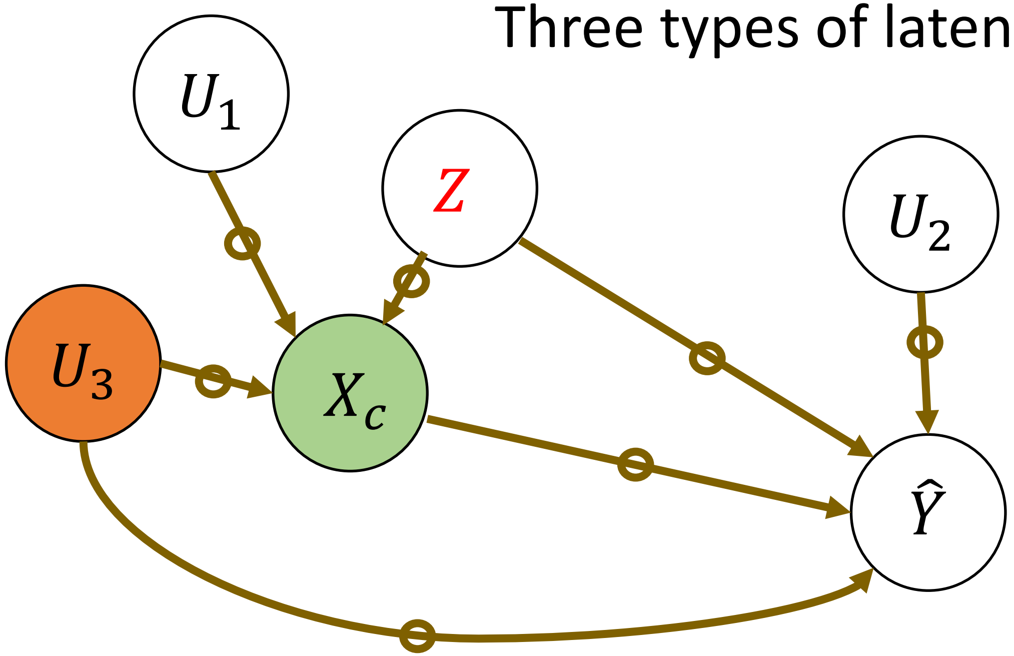


$\mathrm{I}\left((U_a, Z); (\hat{Y}, U_b) | X_c\right)$ is zero

Some intuition on our proposed measure from causality

More generally

Three types of latent source variables U



$\mathrm{I}\left((U_a, Z); (\hat{Y}, U_b)|X_c\right)$ may non-zero for all partitions $U_a = U \backslash U_b$

Non-exempt disparity or Misleading dependencies?

$$\min_{U_a} Uniq\left((U_a, Z):(\hat{Y}, U_b)|X_c\right) \leq \min_{U_a} I\left((U_a, Z);(\hat{Y}, U_b)|X_c\right)$$

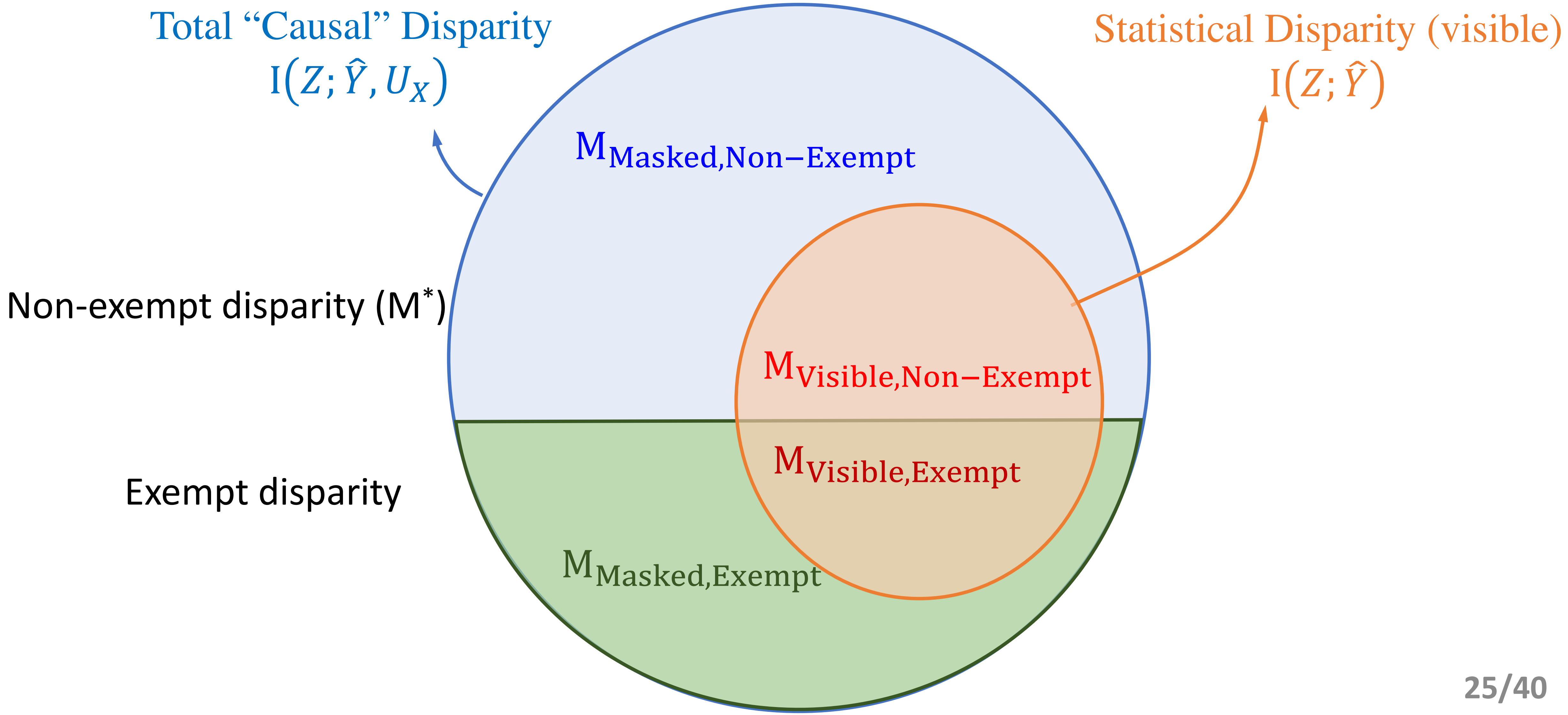
Proposed Measure

"Misleading"

for any set $U_a = U \backslash U_b$

<u>Non-negative</u> decomposition of total "causal" disparity

Theorem 2 (pictorially illustrated)



Observational measures of non-exempt disparity

Theorem: No purely observational measure of non-exempt disparity can satisfy all six desirable properties.

With partial knowledge/assumption about the causal relationships, they may correctly quantify **non-exempt disparity**

Candidate 1:
M= $I(Z;\hat{Y}|X_c)$

Candidate 2:
M= $Uniq(Z:\hat{Y}|X_c)$

Candidate 3:
M= $I(Z;\hat{Y}|X_c,X')$

Case Studies: Artificial Data & Real Data

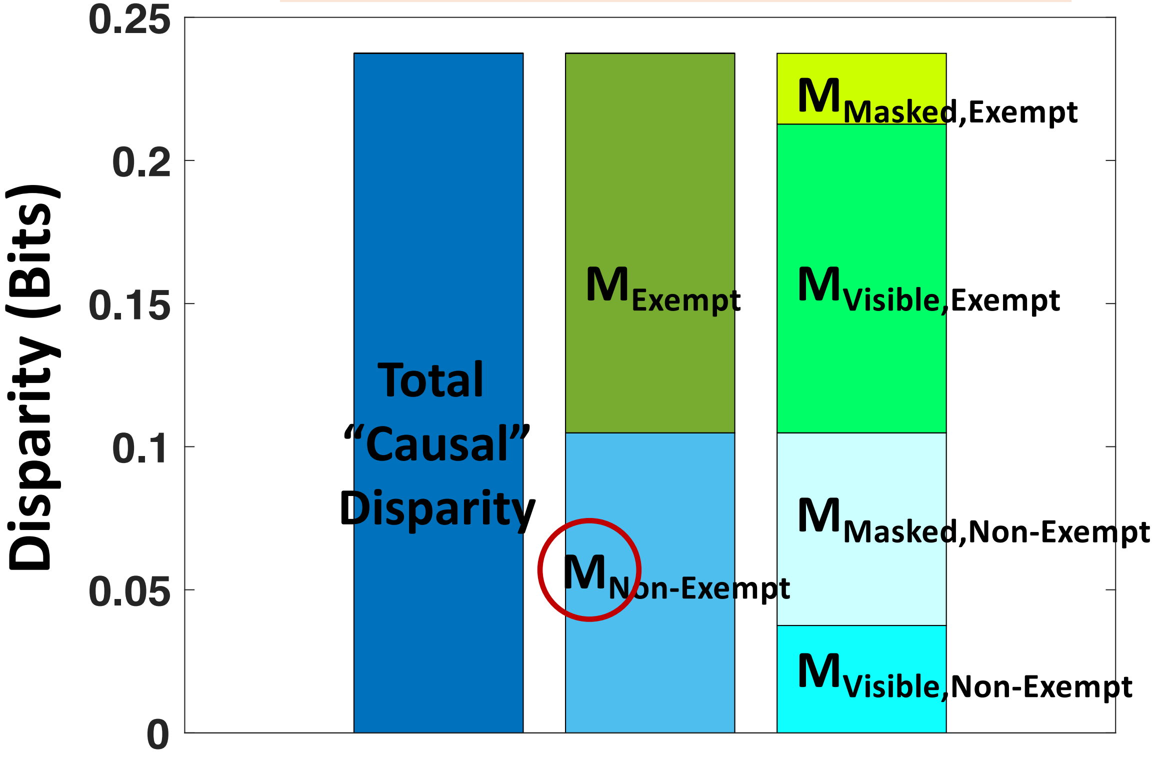
Auditing: Compute causal/observational measures on pre-trained models

Training: $\min_{h(.)} \mathrm{Loss}\left(Y, \hat{Y}\right) + \lambda \underbrace{\mathrm{M}}_{\substack{\textbf{non-exempt disparity} \\ \textbf{(Observational)}}}$ where $\hat{Y} = h(X)$

Simulation: Four types of disparities present

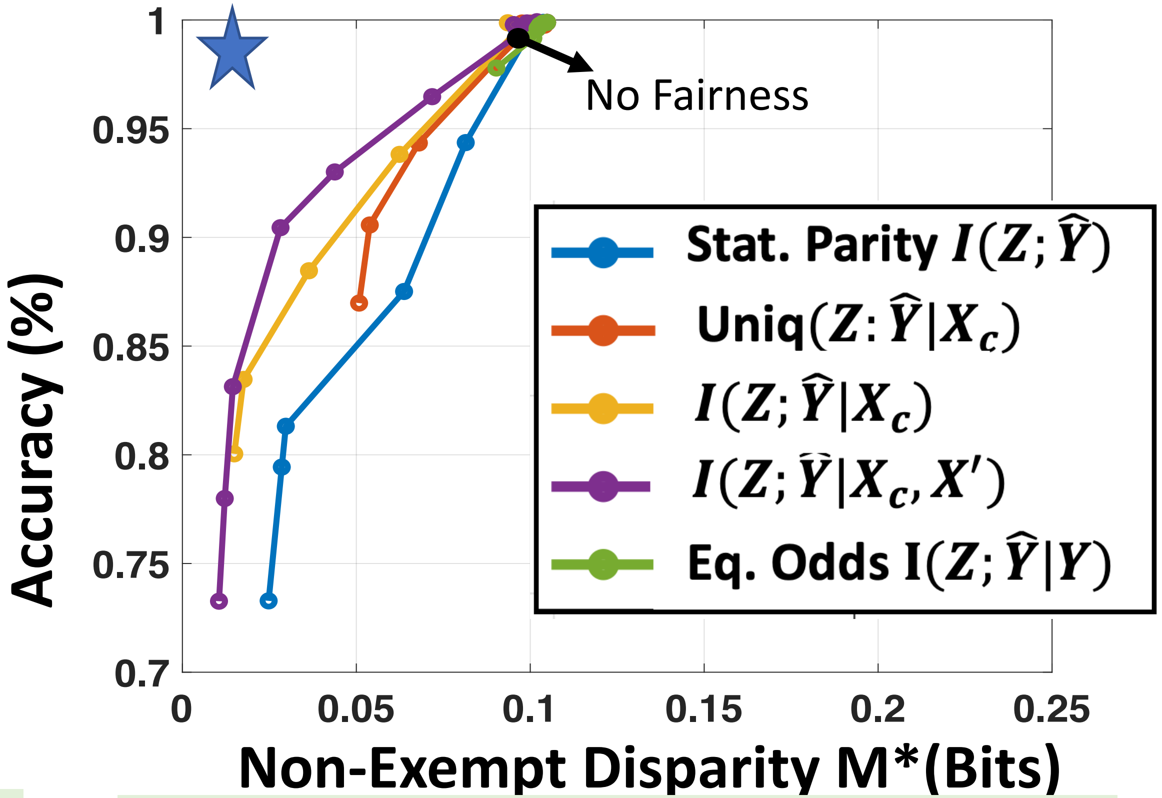
Critical (Writing Sample: $Z + U_1$), General (Browsing History: $Z + U_2$, Proximity: U_3)
Historic True Labels based on equally weighted combination of these features

Auditing a model trained
with no fairness regularizer



Output closely resembles Historic True Labels
- Four types of disparities present

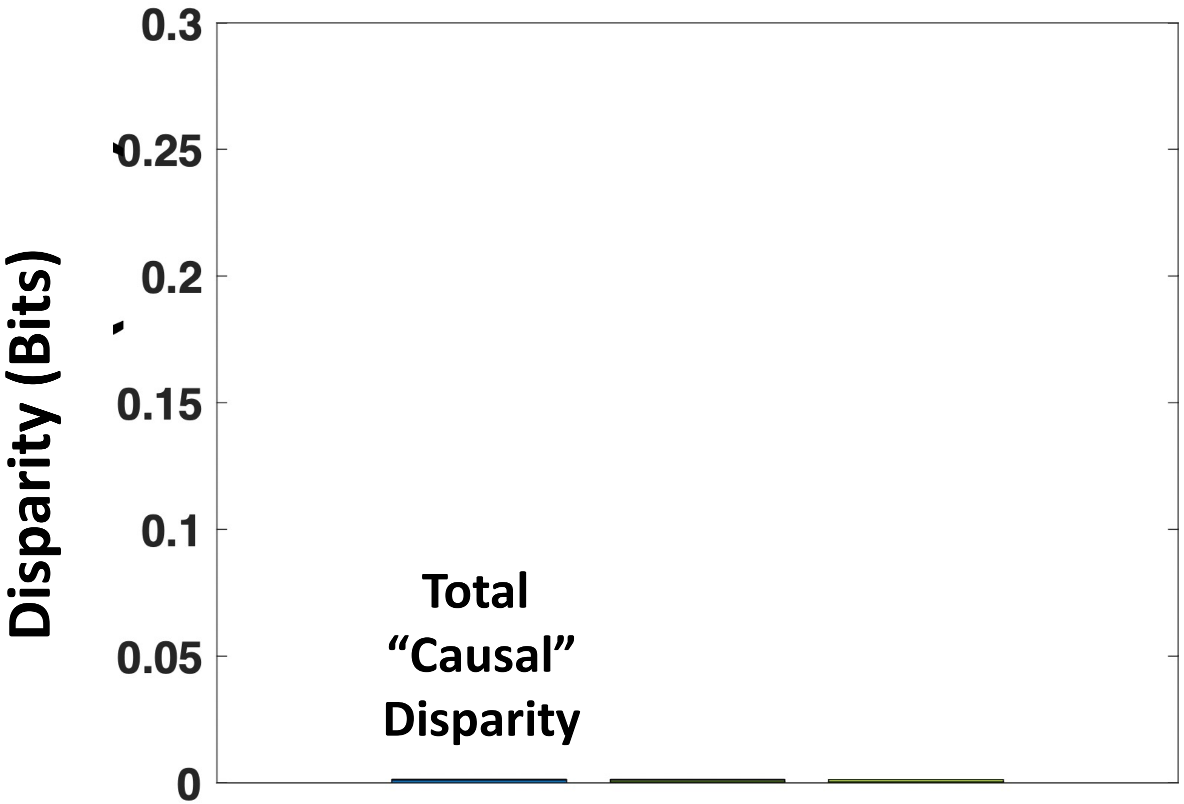
Training models with
different **observational** regularizers



Proposed observational measures
attain better tradeoff

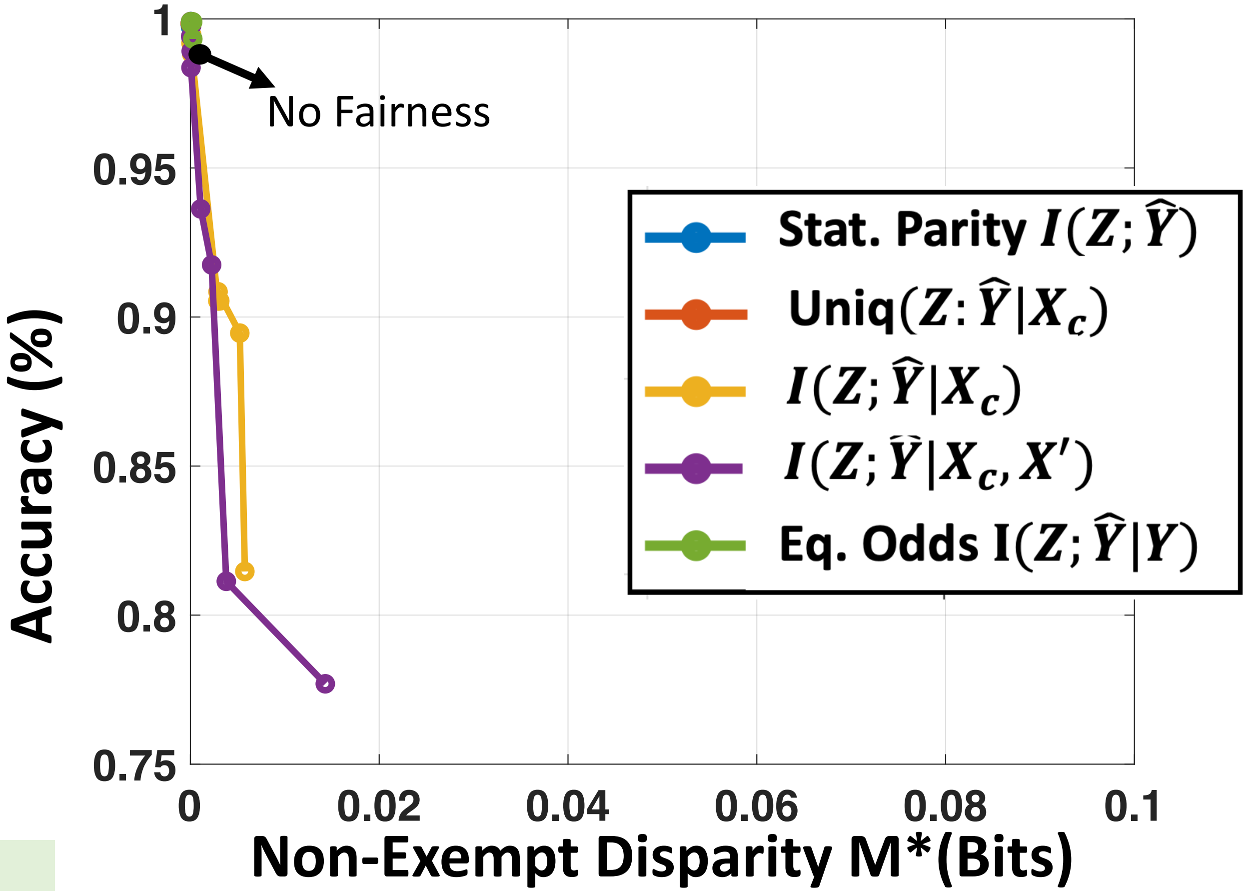
Simulation: No "causal" disparity

Auditing a model trained
with no fairness regularizer



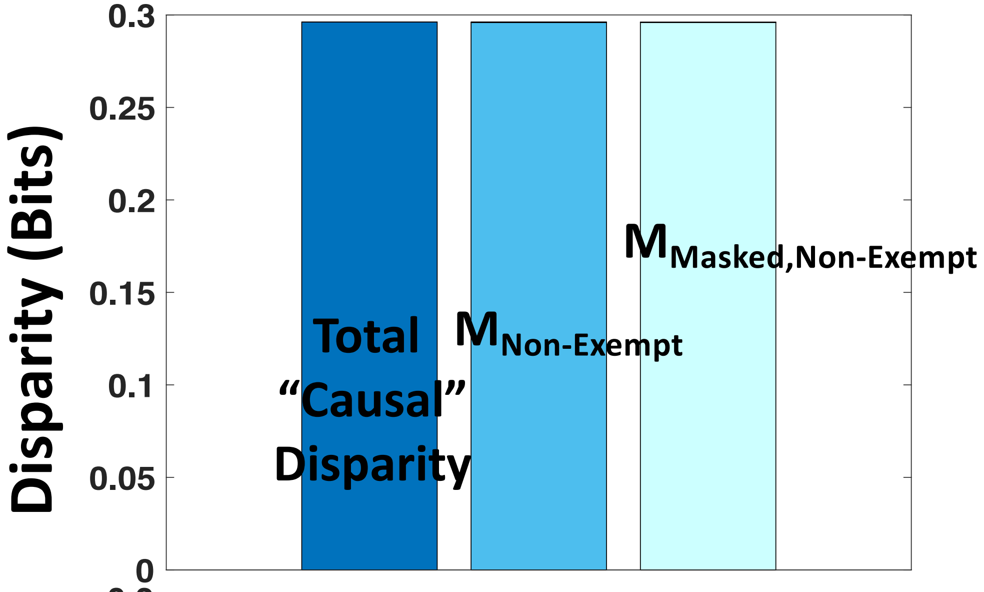
Historic True Labels have no disparity at all
- output with no fairness has negligible disparity

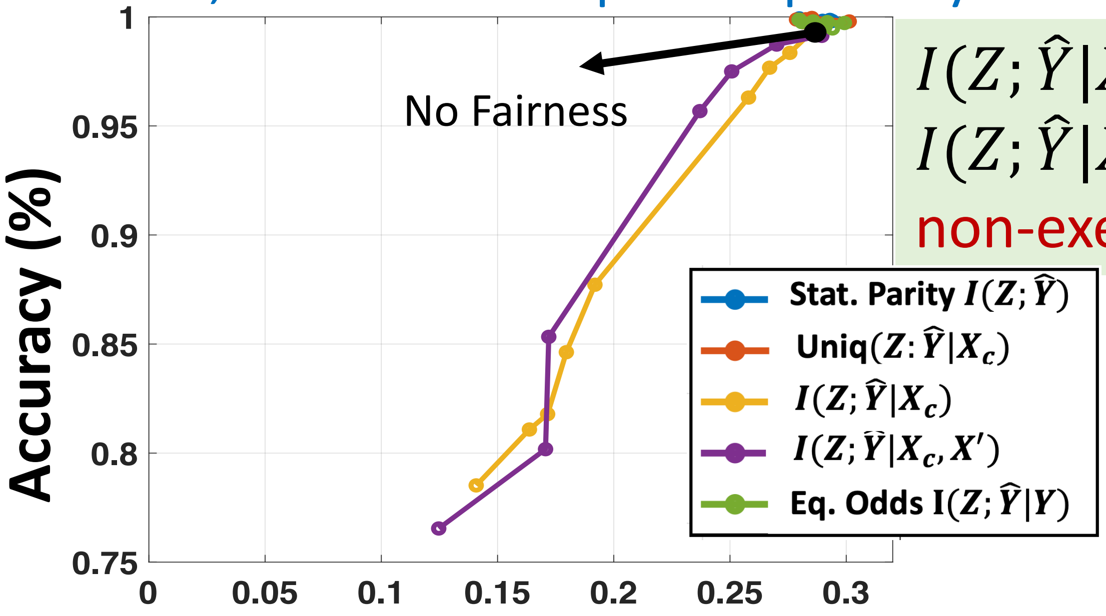
Training models with
different **observational** regularizers



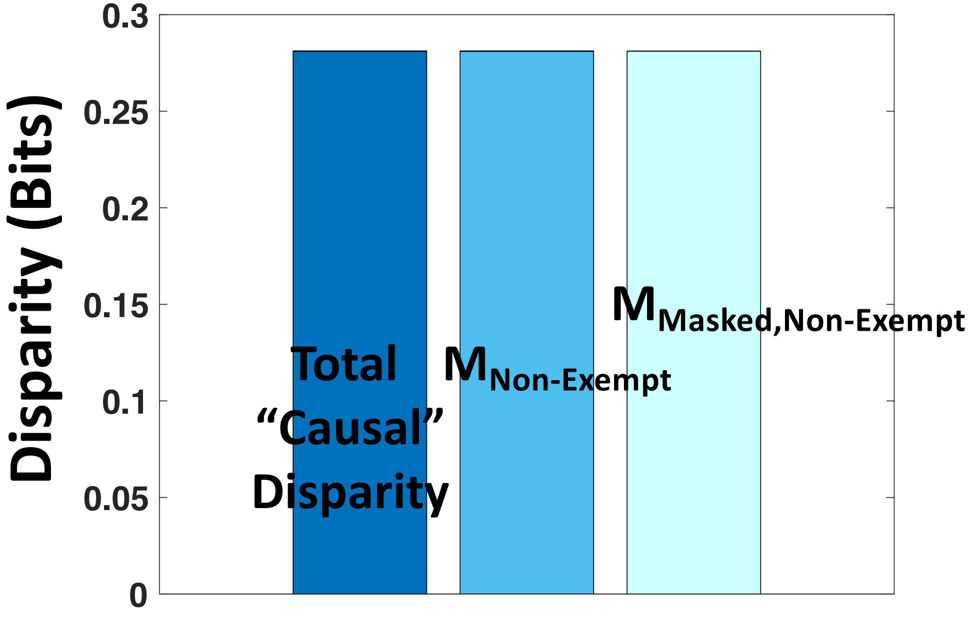
Conditioning falsely detects disparity

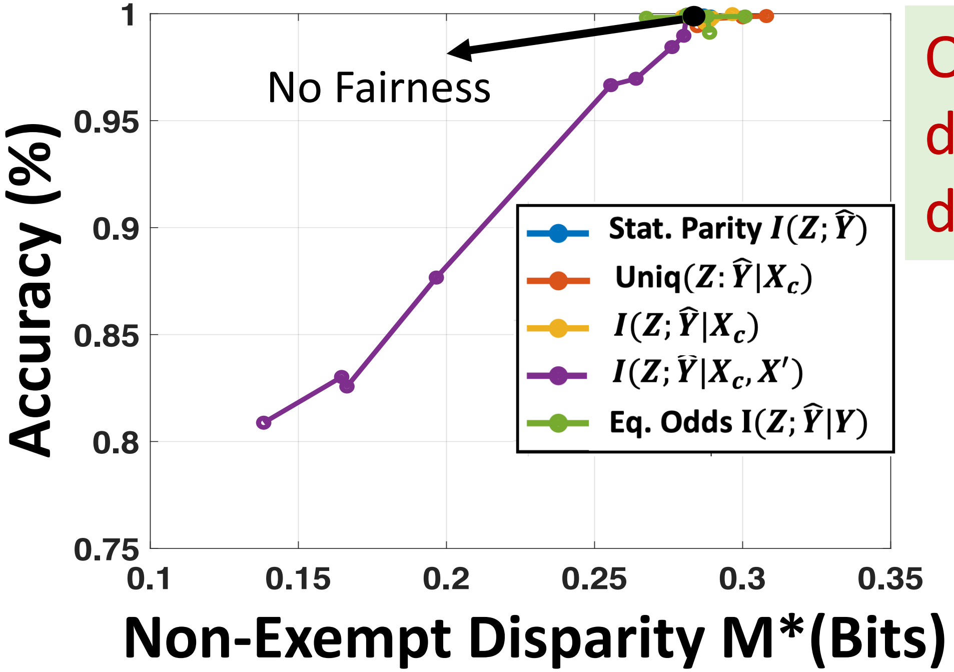
Simulation: Masked, non-exempt disparity





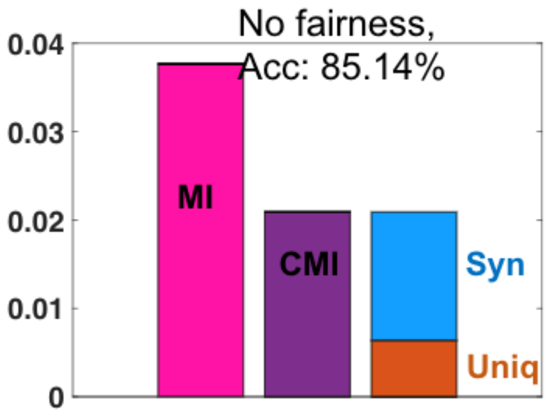
$I(Z;\hat{Y}|X_c)$ and $I(Z;\hat{Y}|X_c,X')$ detect non-exempt disparity



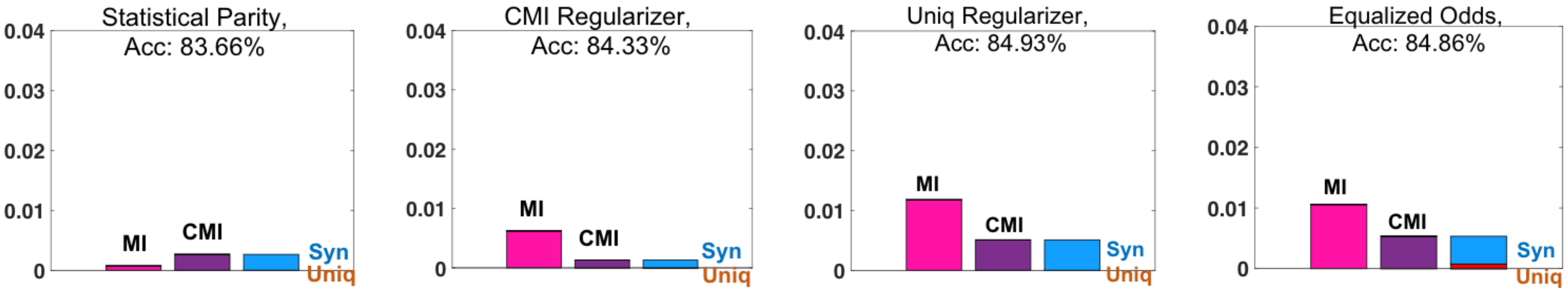


Only $I(Z;\hat{Y}|X_c,X')$ detects non-exempt disparity

Experiment on real data: Causal relationships are not known



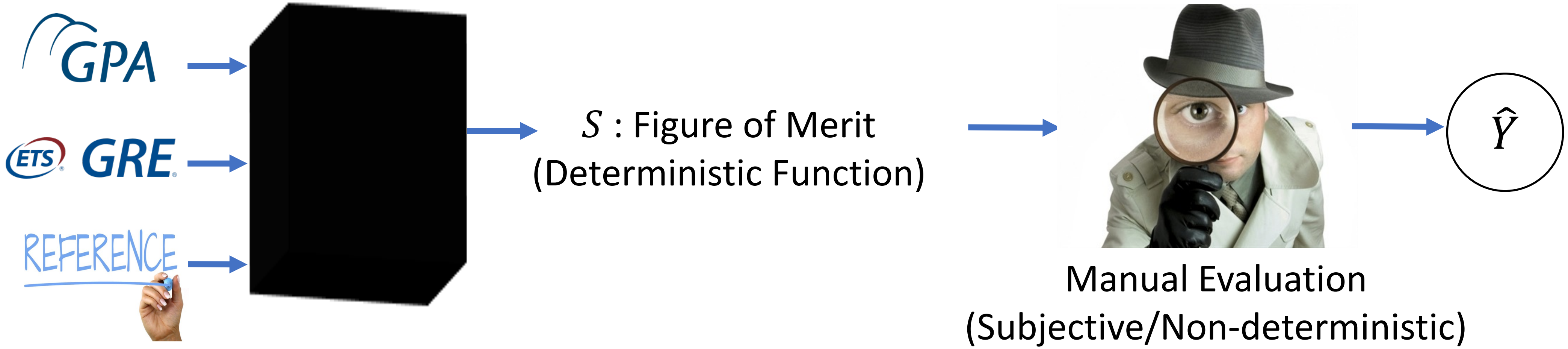
MI: $I(Z;\hat{Y})$
CMI: $I(Z;\hat{Y}|X_c)$
Uniq: $Uniq(Z;\hat{Y}|X_c)$
Syn: CMI-Uniq

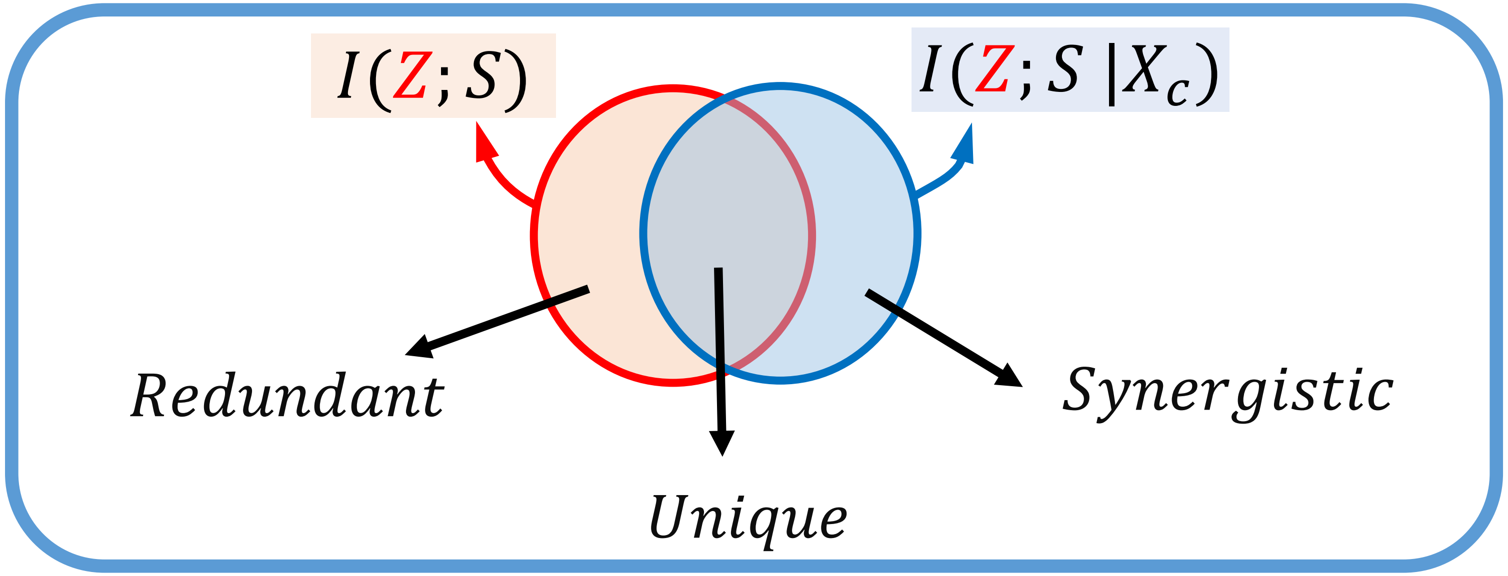


Experiments on Adult Dataset:
Observational relaxations can be used for auditing or training

Similar experiments on German Credit Data

Experiments on CMU ECE Graduate Admissions Dataset as part of ECE Diversity Committee





CMU ECE Graduate Admissions Data from Fall’15 to Fall’18

A summary of our contributions before we move on …

- Systematic approach to find a measure of **non-exempt disparity**
  - Causality + Partial-Information-Decomposition-based measure
  - Observational relaxations
- Conditional Mutual Information $I(Z;\hat{Y}|X_c)$
  - Can falsely detect disparity even if causally fair
- Unique Information $Uniq\left(Z:\hat{Y}|X_c\right)$
  - Doesn't falsely detect disparity but can miss masking
- Preliminary analysis on real data
  - Future Work: Improved Estimators

Broader conversations that this work opens:

- Interpretation/reform of laws for algorithmic hiring
- Essential to collaborate with lawyers/social scientists/minorities

Outline

How to identify/explain the sources of disparity in machine learning models?

Find a measure of non-exempt disparity

[AAAI 2020; IEEE Trans. Info Theory 2021]

Beyond Fairness: Application to Social Media & Filter Bubbles

[BIAS@ECIR 2021]

Perspectives on Accuracy-Fairness Tradeoffs

[ICML 2020] [NeurIPS 2021]

Connections with Explainability

[Workshop@AAAI 2022]

Beyond Fairness: Application to Social Media & Filter Bubbles

Can we debias ***Filter Bubbles*** in social media?

*[Wu, Jiang, **Dutta,** Grover, BIAS@ECIR'21]*

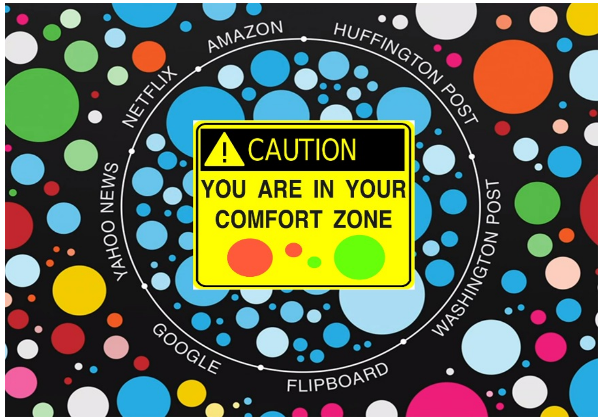
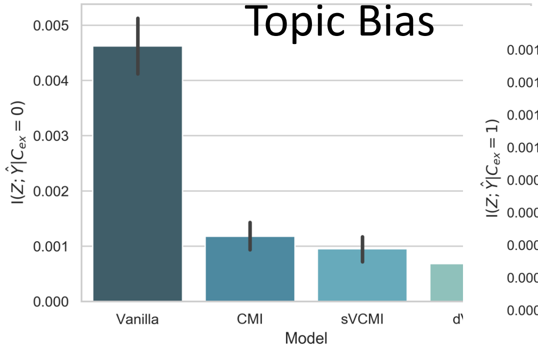


Fig. & Definition: [Pariser'11]

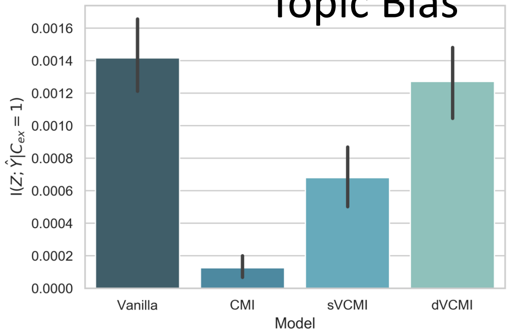
Case Study +
Creation of a new Dataset

Experiments on Artificial Dataset created from Twitter News Sharing User Behavior Dataset

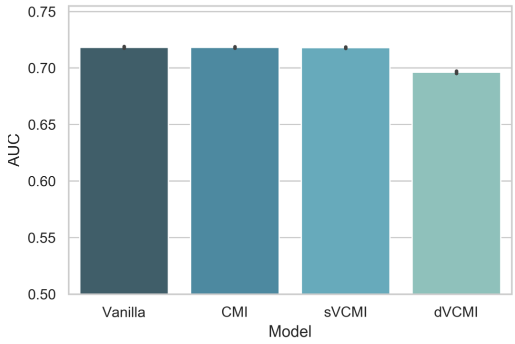
Non-Exempt Topic Bias



Exempt Topic Bias



Performance



News Sharing Behavior of Twitter Users [Brena et al.'19] [Misra'18]

Is there a Tradeoff between Accuracy and Fairness?

Main Contribution:

Quantify Information-Theoretic Limits + Explain They Exist/Don't Exist

*[**Dutta**, Wei, Yueksel, Chen, Liu, Varshney, ICML 2020]*

Key Tool: Chernoff Exponents

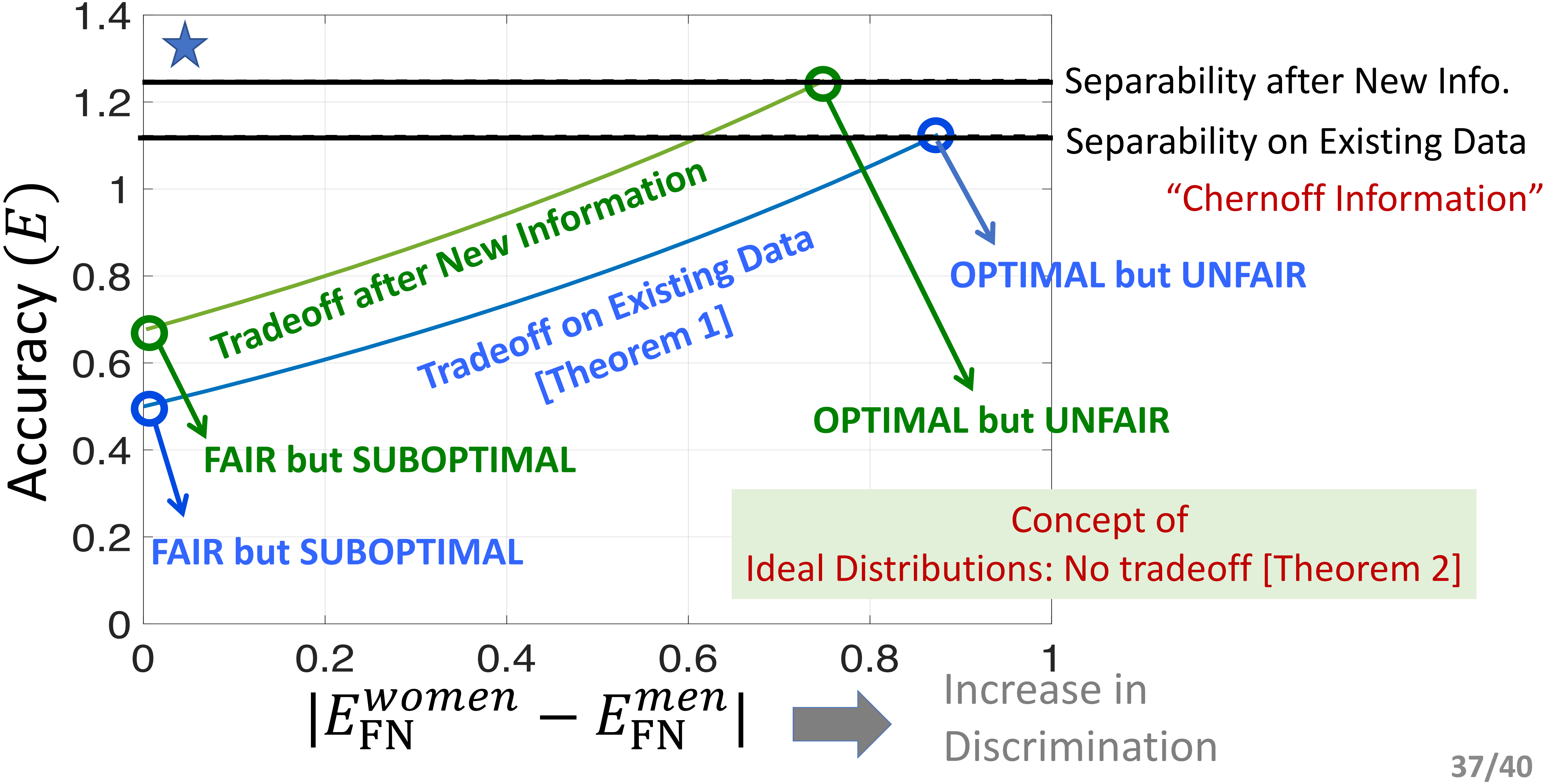
Approximations to the actual error exponents in binary classification

$$P_{FN} \lesssim e^{-E_{FN}} \qquad P_{FP} \lesssim e^{-E_{FP}}$$

Geometric interpretability helps quantify tradeoff between Accuracy and Discrimination in terms of Chernoff Exponents

Related Works: [Menon & Williamson'18][Chen et al.'18][Zhao et al.'19][Sharma et al.'20][Garg et al.'19]

Numerical Computation of Fundamental Limits on the Tradeoff



Looking Forward

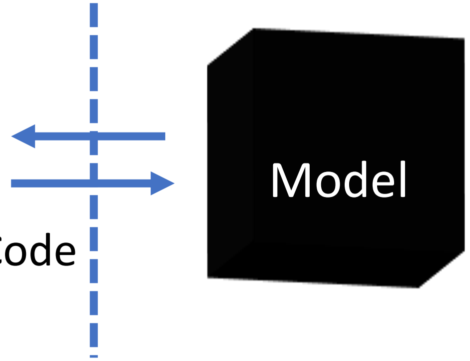
Reliable Machine Learning

Systematic Feature Engineering With Exemptions

Should we even include all features?

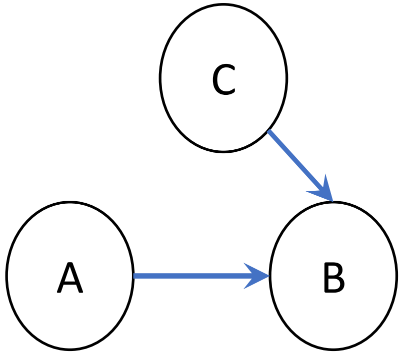
Training Models Under Restricted Access to Certain Features

Restricted Access
To Some Features,
E.g., Gender, Race, Zip Code



Laws can be contradictory [Ricci v. DeStefano'09]
Feature Selection: [Galhotra et al.'20]
Fairness & Privacy: [Mozannar et al.'20][Coston et al. '19]
Epistemic Values & Lived Experiences [Hancox-Li & Kumar'21][Tao & Varshney'21]

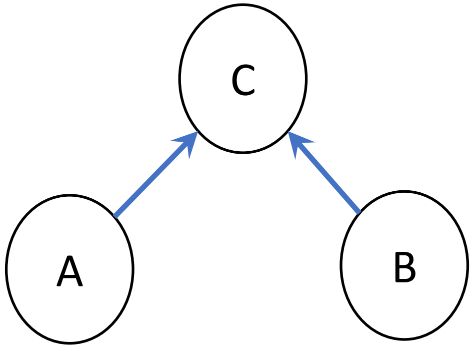
Partial Information Decomposition + Causality



$I(A; B|C) > 0$

$But, Uniq(A: B\backslash C) > 0$

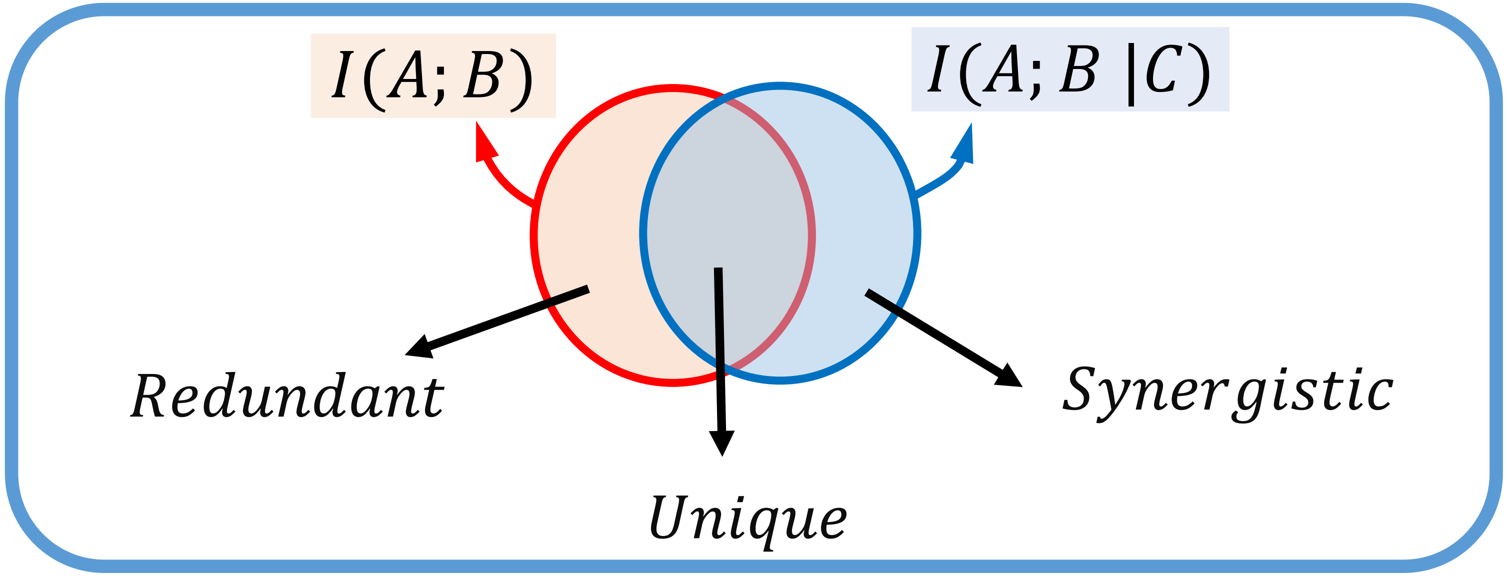
$Syn(A: B, C) = 0$



$I(A; B|C) > 0$

$But, Uniq(A: B\backslash C) = 0$

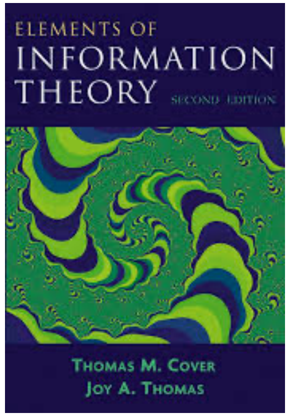
$Syn(A: B, C) > 0$



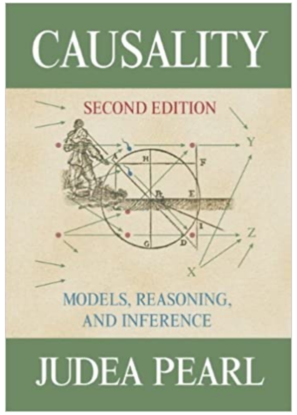
My Research Vision

Connecting with People's Lives

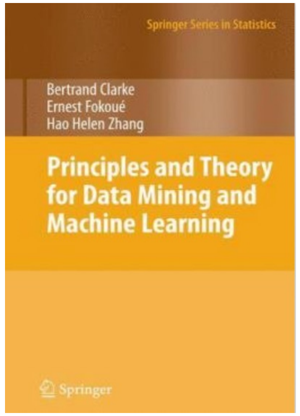
Foundations of Reliable Machine Learning



Information & Coding Theory



Causal Inference



Probability & Statistics

Thank You!

Lawful Hiring

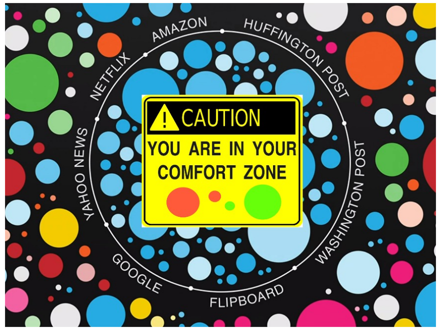
E.g., Design/Audit of Resume Classifier, Ranking, Ads, etc.

Education, Lending

E.g., Explain sources of bias, Recommend interventions, Policy Implications

Social Media & Filter Bubbles

E.g., Political Inclination, Polarization



Healthcare

Robust ML

Federated Learning

Crowdsourcing

(Fairness, Privacy, Reliability)